LBNL-59860 Abs

Novel Bioinformatics Methods for Troubleshooting of Genomic Shotgun Data.

Eugene Goltsman[1], Randal Cox[2], Michael Mazur[3], Alla Lapidus[1], Alex Copeland[1]

[1] Joint Genome Institute, Walnut Creek, CA
[2] Department of Biochemistry and Molecular Genetics; University of Illinois at Chicago, Chicago, IL 60607.
[3] Integrated Genomics, Chicago, IL

End-sequencing of shotgun libraries of small genomic inserts is, by far, the most popular approach to Whole Genome Sequencing (WGS) today. Irregularities in WGS datasets present assembly problems that are expensive and time-consuming to solve, with cloning bias, contamination and long repeats posing the biggest challenges. Shotgun assembly data exhibit well recognizable patterns that follow certain statistical models, and deviations from these models usually stem from flaws and abnormalities in the input data, which, in turn, reflect problems in the cloning protocol, chemistries, or in the DNA being sequenced. We developed several statistical and bioinformatic methods for detecting cloning bias, DNA contamination and high repeat content at early stages of the WGS project. These methods are based on analyses of a) depth of coverage distributions, b) progressive assembly dynamics and c) GC composition distribution of real and simulated shotgun datasets. We identify and describe relationships between coverage (in terms of read depth and number of gaps), and the binomial/Poisson function, and demonstrate ways to routinely identify cloning bias and contamination by relying on these relationships. Differences in GC composition between different genomes, libraries and even plates allowed us to identify cases of suspected contamination by identifying bimodal patterns in the GC distribution in the sequences of a genomic project. Routine automated application is also discussed.

This work was performed under the auspices of the US DOE of Science, Biological and Environmental Research Program, and by the University of California, LLNL under Contract No. W-7405-Eng-48, LBNL under Contract No. DE-AC02-05CH11231 and LANL under Contract No. W-7405-ENG-36.